Regression Methods for Virtual Metrology of Layer Thickness in Chemical Vapor Deposition

Hendrik Purwins, Bernd Barak, Ahmed Nagi, Reiner Engel, Uwe Höckele, Andreas Kyek, Srikanth Cherla, Benjamin Lenz, Günter Pfeifer, and Kurt Weinzierl

***Abstract*—The quality of wafer production in semiconductor manufacturing cannot always be monitored by a costly physical measurement. Instead of measuring a quantity directly, it can be predicted by a regression method (virtual metrology). In this paper, a survey on regression methods is given to predict average silicon nitride cap layer thickness for the plasma-enhanced chemical vapor deposition dual-layer metal passivation stack process. Process and production equipment fault detection and classification data are used as predictor variables. Various variable sets are compared: one most predictive variable alone, the three most predictive variables, an expert selection, and full set. The following regression methods are compared: simple linear regression, multiple linear regression, partial least square regression, and ridge linear regression utilizing the partial least square estimate algorithm, and support vector regression (SVR). On a test set, SVR outperforms the other methods by a large margin, being more robust toward changes in the production conditions. The method performs better on high-dimensional multivariate input data than on the most predictive variables alone. Process expert knowledge used for *a priori* variable selection further enhances the performance slightly. The results confirm earlier findings that virtual metrology can benefit from the robustness of SVR, an adaptive generic method that performs well even if no process knowledge is applied. However, the integration of process expertise into the method improves the performance once more.**

Manuscript received April 7, 2012; accepted May 19, 2012. Date of publication July 26, 2013; date of current version January 17, 2014. Recommended by Technical Editor S. Fatikow. This work was supported in part by the European Union ENIAC project IMPROVE (Implementing Manufacturing science solutions to increase equipment productivity and fab performance) under Grant 120005 and in part by the German BMBF (Federal Ministry of Education and Research) under Contract 13N10417. The work of H. Purwins was supported in part by the German BMBF under Grant BFNT, Fkz 01GQ0850.

H. Purwins was with the Neurotechnology Group, Berlin Institute of Technology, 10587 Berlin, Germany, and also with PMC Technologies GmbH, 48151 Münster, Germany. He is now with the Sound and Music Computing Group, Aalborg University Copenhagen, 2450 Copenhagen, Denmark (e-mail: hpurwins@gmail.com).

B. Barak, R. Engel, U. Höckele, A. Kyek, B. Lenz, G. Pfeifer, and K. Weinzierl are with Advanced Process Control, Infineon Technologies AG, 93049 Regensburg, Germany (e-mail: bernd.barak@infineon.com; reiner.engel@infineon.com; uwe.hoeckele@infineon.com; andreas.kyek@infineon.com; benjamin.lenz@infineon.com; guenter.pfeifer@infineon.com; kurt.weinzierl1@infineon.com).

A. Nagi was with PMC Technologies GmbH, 48151 Münster, Germany, and also with the Music Technology Group, Universitat Pompeu Fabra, 08018 Barcelona, Spain. He is now with R^2 Financial, Toronto, ON M5V 1P9, Canada (e-mail: ahmed.nagi@gmail.com).

S. Cherla was with PMC Technologies GmbH, 48151 Münster, Germany, and also with the Music Technology Group, Universitat Pompeu Fabra, 08018 Barcelona, Spain. He is now with the School of Informatics, City University, London, EC1V0HB, U.K. (e-mail: cherla.srikanth@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMECH.2013.2273435

***Index Terms*—Regression analysis, semiconductor device measurement, silicon semiconductors, virtual metrology.**

I. Introduction

IN a fab, a plant that manufactures semiconductor devices, starting with a uniformly doped bare silicon wafer, the fabrication of integrated circuits needs hundreds of sequential process steps which can be assorted into seven main process areas: *lithography, etching, deposition, chemical mechanical planarization, oxidation, ion implantation,* and *diffusion* [1]. In this paper, we will focus on the deposition step, where a multitude of layers of different materials are deposited onto the production wafers. In order to increase the efficiency of these processes, an advanced fab is required to have online quality monitoring tools. In current practice, process quality is regularly monitored by the sampling of production wafers. This approach assumes that the process quality of production equipment does not change abruptly and that the measurement result of the sampled wafers is a good representative of the actual production quality [2]. This practice may not allow to timely detect equipment process shifts and drifts happening between the scheduled measurements. As a consequence, the quality of the produced wafers may degrade and the production cycle time as well as the cost may increase.

In virtual metrology (VM), the quality of a wafer is predicted based on the process and production equipment data, without physically conducting costly quality measurements [3]. If VM predicts an abnormal equipment state outside the specification limits, it could trigger a stop of the production equipment. The VM can also be used to compensate for minor shifts and drifts of the process data, causing a reconfiguration of the control parameter of the equipment through a run-to-run (R2R) controller [4].

In the deposition steps, chemical vapor deposition (CVD) is applied. CVD is a chemical reaction of a gas mixture at the surface of the wafer taking place at high temperatures. In order to avoid the need of high temperature, in plasma-enhanced chemical vapor deposition (PECVD), the chemical reaction is enhanced by means of electrical fields at radio frequency [5]. An important aspect of this technique is the well defined and reproducible composition and thickness of the deposited film, achievable with reasonable effort by control of the significant process parameters [1]. The *PECVD metal passivation process* considered here comprises the primary deposition of a silicon oxide (SiO_2) base layer onto a metal layer stack and the subsequent deposition of a silicon nitride (Si_3N_4) cap layer (see Figs. 1 and 2). We will present approaches how the Si_3N_4 layer thickness (target) can be predicted based on the process and production equipment data (predictor variables).

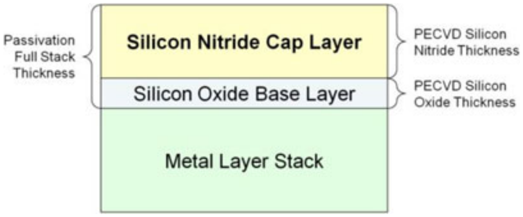
Passivation Full Stack Thickness
Silicon Nitride Cap Layer
PECVD Silicon Nitride Thickness
Silicon Oxide Base Layer
PECVD Silicon Oxide Thickness
Metal Layer Stack

Fig. 1. Metal passivation layer structure: silicon nitride cap layer and silicon oxide base layer, deposited in a PECVD process sequence for passivation of the underlying metal layer stack.

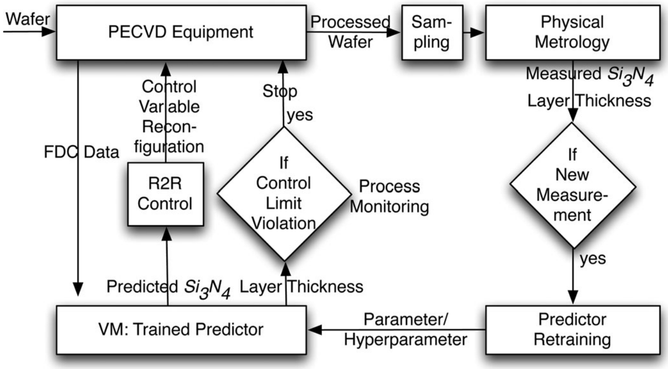

Fig. 2. Given FDC data, the VM predicts the Si_3N_4 layer thickness. These predictions are fed into the run-to-run controller, resulting in the reconfiguration of the control variables of the PECVD equipment. The predictions are also used by the monitoring process to possibly trigger a production stop based on control limit violations. The processed wafers are regularly sampled and subject to physical metrology. In that case, the predictor is retrained and its hyperparameters and parameters are updated.

Relevant previous work comprises [6], in which the authors make use of the Monte Carlo simulation to enhance the design of experiment (DoE) datasets, and model the relation between the input variables and the output variable using a back propagation neural network. In [2], the authors propose to use a radial basis function (RBF) neural network to model the dependence of the output variable on the input variable. However, the models are derived using process parameter data from real production equipment, instead of DoE datasets. Comparing a multiple linear regression (MLR) model (using a stepwise procedure for selecting the input variables) with a multilayer perceptron neural network and a RBF neural network is the focus of a study carried out in [7]. The authors in [8] used the MLR with stepwise selection in order to determine a set of input variables, which are in turn fed into a back propagation and a simple recurrent neural network model to predict the output variable. In [9], the dataset is divided into an in-spec and an out-of-spec dataset, and the classification and regression tree is used to predict when a production wafer will be inside or outside of the defined specification limits. Cheng *et al.* [10] introduced an automatic VM system including a retraining procedure for VM in CVD for the adaptation to different process chambers online, calculating reliance, data similarity, process, and metrology data quality indices. Pampuri *et al.* [11] proposed a hierarchical framework of contexts such as production chambers and production processes, based on a multilevel version of the LASSO, a regularized version of the least-squares problem. Schirru *et al.* [12] compared regularized entropy learning with kernel ridge regression for the prediction of silicon nitride ($\mathrm{Si_3N_4}$) layer thickness also predicting probabilistic uncertainty. For two unspecified etching processes, in [13], variable selection and dimension reduction is used in conjunction with regression (linear and *k*-nearest neighbor regression, regression trees, neural networks, and SVR [14]). On a dataset spanning three months of production, they got the best results with a wrapper feature selection (stepwise linear regression or genetic algorithm with SVR) combined with SVR as a predictor, thereby risking overfitting the data, since all target labels are used twice: by the wrapper algorithm to select the features, and in the end again for evaluation.

As an extension of [15], this paper adds a comparison of SVR to methods based on linear regression and is structured as follows. In Section II, we outline the physical metrology. Section III gives an introduction to the applied regression methods, followed by a description of the performed regression hyperparameter optimization via cross validation (CV) and grid search. In Section IV, we explain the data preprocessing and assess the prediction error for the different methods separately on training and test dataset. Finally, the conclusions are summarized in Section V.

II. Optical Layer Thickness Measurement

The thickness of the silicon nitride layer can be optically measured (physical metrology). In current practice, this expensive procedure is performed on a relatively large number of sampled wafers. To reduce these expenses and to continuously monitor wafer quality, it is desirable to predict this layer thickness via VM instead of actually measuring it. To build such a predictor, we need to train a regression algorithm with a set of actual physical measurements.

For each sampled wafer, the silicon nitride layer thickness is individually measured at several measurement points evenly distributed over the wafer. As an indicator for the quality of each measurement result the goodness of fit (GoF) is used. The mean calculated from these measurements is the average thickness of the silicon nitride cap layer. For the considered range of deposited layer thickness, the relative accuracy of the optical measurement is significantly better than 0.024%. Based on this measured layer thickness, the deposition time for the next lot of wafers of the same design type is calculated by a R2R controller in closed-loop mode.

III. Statistical Methods

Let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})' \in \mathbb{R}^d (i = 1, \ldots, n)$ be the n d-dimensional predictor variables, e.g., consisting of fault detection and classification (FDC) sensor and context variables contained in the historical dataset. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)' \in \mathbb{R}^n$ be the measured variables, in our case, the averaged $\mathrm{Si_3N_4}$ layer thickness assessed by an optical measurement. We present methods to make a prediction (virtual measurement) $\hat{y}$ based on an actual FDC variable input vector $\mathbf{z} \in \mathbb{R}^d$. Some parameters such as the regression coefficients $\mathbf{w}$ of these methods can be learned from the historical dataset, using the *training algorithms* discussed in Section III-A. Other configuration parameters (*hyperparameters*) of the methods can be optimized by systematically evaluating the performance on a set of hyperparameter combinations (grid search, Section III-B).

A. Regression

1) MLR: Let

$$y_i = b + w_1 x_{i1} + w_2 x_{i2} \dots w_d x_{id} + n_i = \underbrace{b + \mathbf{w}'\mathbf{x}_i}_{\hat{y}} + n_i \quad (1)$$

for intercept term b and coefficients $\mathbf{w} = (w_1, \dots, w_d)'$, prediction $\hat{y}$ and noise term n_i. In *Ordinary least-square estimate*, we minimize the empirical risk: $\arg\min_{b,\mathbf{w}} \sum_{i=1}^{n} l(y_i, \hat{y}_i)$, with the quadratic loss function (the squared error)

$$l(y_i, \hat{y}_i) = (y_i - (b + \mathbf{w}'\mathbf{x}_i))^2. \quad (2)$$

Let $\mathbf{X} = (\mathbf{x}_1 - \bar{\mathbf{x}}, \dots, \mathbf{x}_n - \bar{\mathbf{x}})'$, $\mathbf{y} = (y_1 - \bar{y}, \dots, y_n - \bar{y})'$ with sample means $\bar{\mathbf{x}}$ and $\bar{y}$. Then, the coefficient parameters can be estimated by the following expression:

$$\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (3)$$

The intercept is estimated by $\bar{b} = \bar{y} - \hat{\mathbf{w}}'\bar{\mathbf{x}}$.

A new data point $\mathbf{z} \in \mathbb{R}^d$ is centerized yielding an estimated measurement

$$\hat{y} = \bar{b} + \hat{\mathbf{w}}'(\mathbf{z} - \bar{\mathbf{x}}). \quad (4)$$

2) Simple Linear Regression (SLR): This method chooses the component from predictor variable $\mathbf{x}$ that gives the lowest squared error and performs regression only with this single variable [$d = 1$ in (1)].

3) Ridge Linear Regression (RLR): When the columns of $\mathbf{X}$ have an approximate linear dependence, the matrix $\mathbf{X}'\mathbf{X}$ becomes close to singular. *Ridge Regression* [16] addresses this problem of multicollinearity by solving the following expression instead of the one in (3):

$$\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X} + r\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \quad (5)$$

where the ridge parameter r is a hyperparameter and $\mathbf{I}$ is the identity matrix. Small positive values of r improve the conditioning of the problem and reduce the variance of the estimates. In Section IV, RLR will be used after previous partial least squares (PLS) filtering (RLR, cf., Section III-A4, [17]).

4) PLS Estimate: For high-dimensional predictor variables $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$, it can be desirable to reduce the dimension d. In PLS, new latent variables, so-called *scores* $\mathbf{t_i}, \dots, \mathbf{t_g}$ $(g < d)$, are constructed, exploiting correlations between the predictor variables and between the predictor variables and the measured variables $\mathbf{y}$. Thereby, the dimension of the input space is reduced, while at the same time, a minimal amount of relevant information is lost.

According to [17], the PLS1 algorithm consists of the following steps.

1) Initialize $j = 1, \mathbf{X}_1 = \mathbf{X}, \mathbf{y}_1 = \mathbf{y}$.

2) Calculate component $\mathbf{v_j}$, pointing at the direction of the largest covariance between $\mathbf{X}_j$ and $\mathbf{y}_j$:

$$\mathbf{v}_j = \frac{\mathbf{X}_j'\mathbf{y}_j}{\|\mathbf{X}_j'\mathbf{y}_j\|}. \quad (6)$$

3) Calculate the scores:

$$\mathbf{t}_j = \mathbf{X}_j\mathbf{v}_j. \quad (7)$$

4) Do linear regression of the measurements $\mathbf{y}_j$ on the score $\mathbf{t}_j$, yielding the regression coefficient $\hat{c}_j = \mathbf{t}_j'\mathbf{y}_j / \mathbf{t}_j'\mathbf{t}_j$.

5) Perform linear regression of $\mathbf{X}_j$ on $\mathbf{t}_j$, resulting in the coefficient vector $\hat{\mathbf{p}}_j = \mathbf{X}_j'\mathbf{t}_j / \mathbf{t}_j'\mathbf{t}_j$.

6) Get residuals after the regression:

$$\mathbf{X}_{j+1} = \mathbf{X}_j - \hat{\mathbf{p}}_j\mathbf{t}_j' \quad (8)$$

$$\mathbf{y}_{j+1} = \mathbf{y}_j - \hat{c}_j\mathbf{t}_j. \quad (9)$$

7) Stop if $j = g$, otherwise let $j = j + 1$ and return to Step 2.

The algorithm yields the components $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_g)$, the scores $\mathbf{T} = (t_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le g}} = (\mathbf{t}_1, \dots, \mathbf{t}_g)$, and the coefficients $\hat{\mathbf{c}} = (\hat{c}_1, \dots, \hat{c}_g)$ and $\hat{\mathbf{P}} = (\hat{\mathbf{p}}_1, \dots, \hat{\mathbf{p}}_g)$. The orthogonality of the columns $\mathbf{t}_j$ of matrix $\mathbf{T}$ ensures that the multiple regression of $\mathbf{y}_j$ on $\mathbf{T}$ can be done one column $\mathbf{t}_j$ at a time (Step 4, [17]). The centered predictor variables $\mathbf{x}_i - \bar{\mathbf{x}}$ can then be approximated as the linear combinations $\hat{\mathbf{P}}(t_{i1}, \dots, t_{ig})'$ with noise term $\mathbf{n}_i \in \mathbb{R}^d$ as follows:

$$\mathbf{x_i} - \bar{\mathbf{x}} = \hat{\mathbf{P}}(t_{i1}, \dots, t_{ig})' + \mathbf{n}_i. \quad (10)$$

Analogously, the measurements $\mathbf{y}$ can be decomposed in terms of coefficients $\hat{\mathbf{c}}$ with noise term n_i as follows:

$$y_i = \hat{\mathbf{c}}(t_{i1}, \dots, t_{ig})' + n_i. \quad (11)$$

For the prediction $\hat{y}$ of the measurements, the variable vector $\mathbf{z}$ is centerized (4) and its score $\mathbf{t}^* = (t_1, \dots, t_g)'$ is recalculated iterating (6) and (7) g times [17], yielding

$$\hat{y} = \bar{y} + \hat{\mathbf{c}}\mathbf{t}^* \quad (12)$$

for sample average $\bar{y}$. g is a hyperparameter to be optimized.

PLS can be considered as a method based on model order selection and dimension reduction and might be used as an alternative to the expert selection of variables described in Section IV-A4.

5) Support Vector Machine Regression (SVR [14]): Generalizing (1), a (nonlinear) transformation φ is introduced into the regression equation

$$y_i = b + \mathbf{w}'\varphi(\mathbf{x}_i) + n_i = \hat{y}_i + n_i. \quad (13)$$

To estimate b and $\mathbf{w}$, we minimize $\sum_{i=1}^{n} l(y_i, \hat{y}_i)$. Instead of the quadratic loss (2), we use the ε-insensitive loss function

$$l(y_i, \hat{y}_i) = \begin{cases} 0, & \text{if } |y_i - \hat{y}_i| \le \varepsilon \\ |y_i - \hat{y}_i| - \varepsilon, & \text{else .} \end{cases} \quad (14)$$

Thereby, data points within an ε-*tube* around the estimate $\hat{y}_i$ are not considered for constructing the predictor.

Slack variables are defined as

$$\xi_i := \min(y_i - \hat{y}_i - \varepsilon, 0) \quad (15)$$

$$\xi_i^* := \min(-y_i + \hat{y}_i - \varepsilon, 0). \quad (16)$$

In SVR, the aim is to find the coefficients $\mathbf{w}$ to minimize

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\xi_i^*\right) \tag{17}$$

under the constraints

$$\begin{aligned} y_i - \mathbf{w}'\varphi(\mathbf{x}_i) + b &\le \varepsilon + \xi_i \\ \mathbf{w}'\varphi(\mathbf{x}_i) - b - y_i &\le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* &\ge 0. \end{aligned}$$

C trades off the flatness of $\hat{y}$ versus the prediction error. This optimization problem is solved via the *dual maximization problem* [14].

$$L = -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\varphi(\mathbf{x}_i)'\varphi(\mathbf{x}_j) \tag{18}$$

$$-\varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) \tag{19}$$

under the conditions: $0 = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*), 0 \le \alpha_i, \alpha_i^* \le C$. We yield $\hat{y} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\varphi(\mathbf{x}_i)'\varphi(\mathbf{x}) + b$. In this equation, φ only occurs pairwise. Such a pair can be substituted by a kernel function, typically a (Gaussian) RBF with flatness hyperparameter γ as follows:

$$\varphi(\mathbf{x}_i)'\varphi(\mathbf{x}) = k(\mathbf{x}_i, \mathbf{x}) = e^{-\gamma\|\mathbf{x}_i - \mathbf{x}\|^2}. \tag{20}$$

The SVR has three hyperparameters to be optimized: γ, ϵ, and C.

B. Hyperparameter Optimization by Grid Search

The previous section provided algorithms that learn the regression coefficients $\mathbf{w}$ and $\mathbf{c}$ from the training data. However, some regression methods are further specified by additional hyperparameters, some of which are set based on *a priori* experience. Others are systematically optimized, using the *grid search* procedure explained in the sequel. Whereas MLR and SLR do not have any hyperparameters, RLR has the hyperparameters r (*ridge parameter*) and g (*number of components for PLS*). From s values for r, equally spaced on the logarithmic scale, and t equally spaced values for g, the cross product (*grid*) of all $s \cdot t$ combinations of (r, g) pairs is built, i.e., the grid created by $r = 10^{-10}, 10^{-9}, 10^{-8}, \ldots, 10^{5}$ and $g = 1, 2, 3, \ldots, \lfloor\frac{2 \cdot d}{3}\rfloor$, d being the dimension of the predictor variable vector $\mathbf{x}$. The data are divided into a training subset and an evaluation subset. On the training subset, for each combination of r and g, the two-fold CV (see [18]) root-mean-square-error (RMSE) is calculated, measuring the deviation of the predicted from the measured layer thickness. Around the combination with minimal RMSE, a smaller $3 \cdot 3$ grid of $r \cdot g$ combinations is evaluated with the ten-fold CV RMSE. If the grid point with minimal RMSE lies on the border, another $3 \cdot 3$ grid is centered around there. If necessary, this is iterated up to three grid extensions, yielding an optimal hyperparameter combination. This procedure is performed for ten training/evaluation partitions (ten-fold CV) and repeated five times with different randomization of the order of the data. The best hyperparameter combination is the one that occurs most often across different data randomizations. The latter is then used for the prediction ($\hat{y}$) on the test set (points $\mathbf{z}$), for which the average RMSE and the standard deviation (std) are calculated.

The optimal g for RLR is also used for PLS. For SVR, $\epsilon = 0.1$ in (14) is empirically chosen and the grid is spanned by $C = 10^{-1}, 10^{-0.5}, 10^{0}, \ldots 10^{3}$ [in (17)] and $\gamma = 10^{-3}, 10^{-2.5}, 10^{0}, \ldots 10^{2}$ [in (20)]. The *WEKA* [19] toolkit is used for implementation.

IV. Results

In order to find the best VM method to predict the $\mathrm{Si_3N_4}$ layer thickness based on sensor and context variables, the regression methods from Section III with different parameterizations are compared through validation and evaluation on a historical dataset recorded from the production (see Section I) and the metrology equipment (see Section II). The historical dataset is split into two portions. The first portion of sensor and context variables consists of 450 variable instances (wafers) recorded over a period of nine months and is preprocessed (see Section IV-A) yielding a *training set* on which various (parameter) settings for the regression methods are compared (see Section IV-B1). The best settings are then applied to a *test set* (see Section IV-B2) from the second portion of the historical dataset, recorded during a period of five months, starting five months after the last variable instance of the training set.

A. Data Preprocessing

First, from more than 150 FDC context and sensor variables in the historical dataset, the CVD experts selected a variable subset according to its relevance for the $\mathrm{Si_3N_4}$ metal passivation step considered here. In addition, only wafers for a particular product technology type were chosen. Further variables are removed if statistically irrelevant (see Section IV-A1). Wafer instances with missing or inconsistent variable values or outside certain specified limits are discarded. In Section IV-A2, it is investigated that whether the two remaining context variables, the *basic* (wafer) *design type* and the *process chamber*, have a statistically significant impact on the $\mathrm{Si_3N_4}$ thickness, using *analysis of variance (ANOVA)*. Based on the latter analysis, instances are only kept if their combination of context variables occurs frequently enough to be used for statistical inference (see Section IV-A3). From the remaining variable subset with 41 variables, the experts selected two further variable subsets containing 17 and 3 variables, used to compare the performance of the different regression algorithms (see Section IV-A4).

1) Statistical Filtering: After the initial expert variable preselection, variables and wafer instances are further reduced according to the following sequence of filtering rules:

a) removal of variables with entirely or predominantly missing values;
b) removal of variables with constant values;
c) removal of context variables being redundant or without process relevance;
d) removal of instances with missing variable values;

e) removal of instances with particular variable (*mean GoF, minimum GoF, data gap count*) values outside specified limits (cf., Section II);

f) removal of the variables used for threshold filtering in the previous step;

g) removal of instances with inconsistent variable values.

The resulting FDC dataset used for training comprises 414 instances and two context variables: 1) the *basic design type*, characterizing the physical properties of the wafer and 2) the *process chamber* index, indicating in which of the three process chambers of the recorded PECVD equipment the wafer has been processed.

2) ANOVA of Context Variables: We utilize the one-way ANOVA to assess whether the measured average Si_3N_4 layer thickness has the same mean for wafer groups of different context variables, i.e., *basic design type* and *process chamber*. This indicates, whether the bias (intercept) term for regression modeling has to depend on these context variables.

a) Bias by Basic Design Type: For the five most frequent *basic design types*, ANOVA returns a p-value of 2.8×10^{-10}, indicating that the average Si_3N_4 thickness corresponding to the different *basic design types* do not have a common mean. Thus, the bias term b in the regression model (1), has to depend on the *basic design type of the processed wafer*.

b) Bias by Process Chamber: ANOVA for three process chambers results in an even higher p-value than for the *basic design types*, i.e., the probability of the null hypothesis to be 2.4×10^{-12}. Therefore, we build separate regression models for each chamber, taking also into account that statistical process control (SPC) in the manufacturing line is performed specifically for each process chamber.

3) Selecting Frequent Context: To exemplify this approach, in our study, we only consider wafer instances from the process chamber with maximum number of instances. In order to base prediction modeling on an overall statistically significant number of wafer instances, only *basic design types* with an occurrence of more than seven instances in the FDC dataset are considered, i.e., five different *basic design types* with a frequency ranging from 8 to 30. Thus, the number of remaining instances available for statistical analysis totals to 98. Subsequently, two more variables are removed that are constant across the reduced set of remaining instances. In order to utilize the remaining context variable *basic design type* as a numerical predictor variable, the nominal variable *basic design type* is converted into a five-dimensional binary vector, indicating which of the five *basic design types* occurs. Additionally, the values of all parameters in the FDC dataset are mapped to the range $[0, 1]$.

4) Compared Variable Sets: Finally, 41 normalized predictor variables are available for application of the different regression methods. We will call this variable set *filtered full variable set (FF)*. It will be compared with other variable sets. From FF, CVD process experts again selected a subset of variables, using more strict criteria, resulting in a variable set called *expert selected FDC variable set (ES)*. It consists of 17 sensor variables and the context *basic design type* (see Table I). From this set, the FDC experts finally selected the three most important variables: The *recipe set point for wafer deposition time from R2R controller, the temperature of processed wafer (mean), as well as the context basic design type of processed wafer*, thus defining the TTB variable set.

TABLE I
EXPERT SELECTED (ES) VARIABLE SET AND SI_3N_4 TARGET, INCLUDING FORWARDED AND REFLECTED RADIO FREQUENCY (RF) POWER, WITH RELATIVE STD (STD DIVIDED BY MEAN): (CF., GRILL [5])

Predictor Variables	Relative Std
Basic design type of processed wafer (5 binary ind.)	
Process chamber pressure (M,Std)	0.0213/0.1818
Process chamber pressure control valve position (M,Std)	0.0093/0.1758
Nitrogen (N_2) gas flow into proc. chamber (M,Std)	0.0002/0.1667
Monosilan (SiH_4) gas flow into proc. chamber (M,Std)	0.0002/0.0959
RF-power forw. into process chamber > limit (M,Std)	0.0148/0.1967
RF-power refl. from process chamber > limit (M,Std)	0.0745/0.2491
Norm. dev. of refl. RF-power (M) from batch median	3.0394
Temperature of processed wafer (Std,M)	0.0005/0.0904
Recipe set point wafer deposition time from R2R-ctrl.	0.0055
Count processed wafers since proc. chamber wet clean	0.1617
Target: Silicon Nitride layer thickness (M)	0.0087

M: Mean / Std: Standard Deviation.

TABLE II
TRAINING SET PREDICTION OF AVERAGE PECVD SILICON NITRIDE LAYER THICKNESS WITH SLR, MLR, PLR, RLR, AND SVR BASED ON THE TTB, ES, AND FF VARIABLE SETS

Variable Set	Method	CV Rel. RMSE	Rel. Std
Deposition Time	SLR	0.895	0.019
TTB	MLR	1.524	2.540
	PLR	0.391	0.004
	RLR	0.397	0.010
	SVR	**0.374**	0.004
ES	MLR	0.348	0.014
	PLR	**0.322**	0.010
	RLR	0.323	0.012
	SVR	0.339	0.013
FF	PLR	0.371	0.009
	RLR	0.372	0.013
	SVR	**0.344**	0.010
Uncond. VM Acc. Limit		0.507	

Average relative CV RMSE and Relative std are given over five randomizations and ten-fold CV.

5) Test Set: Filtered in the same way as the training set, the test set contains 39 instances for the same variable sets (FF, ES, TTB).

B. Prediction Results

For the training set and the three variable sets (TTB, ES, FF), different regression algorithms and their hyperparameter settings are evaluated (Section IV-B1). This yields the best performing variable set and the optimized hyperparameters for each regression method. These settings are then applied to the test set (see Section IV-B2). In comparison to the validation error on the training set, the *test error* is a more reliable estimate of the performance of the VM algorithms in terms of future productive online application [18].

1) CV Training Error: On the training set, the ten-fold CV performance over five randomizations of the data partitioning is averaged for the five methods, SLR, MLR, PLR, RLR, and SVR, using the TTB, ES, and FF variable sets (see Table II). The performance is evaluated using the average *relative RMSE* being the average RMSE divided by the std of the target variable (Si_3Ni_4 layer thickness). The *relative std* is the RMSE std across the five randomizations divided by the std of the silicon nitride layer thickness. As a reference, we also indicate the *relative unconditional VM accuracy limit* based on the process tolerance specification for the considered product technology type divided by the target silicon nitride layer thickness.

a) Prediction on the Single Most Predictive Variable: SLR determines the most predictive variable, i.e., in our dataset, the *wafer deposition time*, and predicts the layer thickness based on that variable. The CV relative RMSE is more than twice as high as for the other methods and variable sets, except for MLR-TTB, indicating that an univariate linear regression model is suboptimal for predicting the layer thickness.

b) Prediction on Deposition Time, Temperature, and Basic Design Type: Adding the variable *wafer temperature* and the context variable *basic design type* to the *deposition time* (TTB) using PLR, RLR, or SVR let the prediction error decrease dramatically. PLR performs only 1.5% better than RLR with the relative std being 0.4% or 1.0% for the respective algorithm. The average RMSE for MLR is almost four times as high as for PLR and RLR. In this case, the performance depends strongly on the randomization of the training/test partitions, indicated by a relative std of 2.54. The results of MLR without collinearity analysis and feature selection are very sensitive to different randomizations of the dataset, and give a highly unstable predictor. An informal comparison indicates that MLR with a variety of methods for feature selection and for elimination of collinear attributes performs in the same order as PLR and RLR. On the TTB variable set, SVR performs slightly better than the other methods.

c) Prediction on ES Variable Set: On the ES variable set, PLR and RLR perform best and almost equally well with an average relative RMSE of 0.322 (std: 1.0%) and 0.323 (std: 1.2%), respectively. Compared to the TTB results, the performance of PLR on ES is improved by 18%. For SVR, the RMSE is a bit higher than for PLR/RLR. On the ES variable set, MLR performs worst by a small margin.

d) Prediction on FF Variable Set: PLR is only 0.3% better than RLR. With respect to the results obtained for the ES variable set, the RMSE performance of RLR as well as PLR is degraded by 15%, but is still slightly better (5–6%) than the results for the TTB variable set. MLR is not applied to the FF variable set since the related computational effort is not feasible. However on FF, SVR performs best, almost as good as on ES.

PLR/RLR on ES gives the best performance on the training set for all combinations of algorithms and variable sets considered in this paper.

2) Prediction on Test Set:

a) Expert Selected Variable Set: As it performed best for the training set, we choose the ES variable set for the prediction on the test set. The RMSE performance of the five algorithms (SLR, MLR, PLR, RLR, and SVR) is compared. For PLR, RLR, and SVR, the optimal hyperparameter combinations have to be determined. We chose the hyperparameter combinations that performed best in the ten-fold CV on the training set in the majority of the five data randomizations. Using these hyperparameters, the algorithms are trained on the training set. The trained models are then applied to predict the average silicon nitride layer thickness on the test set.

TABLE III
TEST SET PREDICTION BASED ON THE ES VARIABLE SET

Method	Relative RMSE
SLR	1.000
MLR	1.894
PLR	0.759
RLR	0.749
SVR	**0.432**

SLR selects *deposition time* as the only variable used for regression.

In Table III, for the different algorithms, the relative RMSE for the test set prediction based on the ES variable set is shown. The performance on the test set is significantly different from the one on the training set. In contrast to the training error, the best method on the ES test set is SVR with a relative RMSE of 0.432, which is only 27% worse than the SVR training error. On the test set, SVR outperforms the other methods by a large margin. The test error of the second best method RLR is 73% higher than the one of SVR and 132% worse than the RLR training error. For the other methods, the increase of the test error with respect to the one of SVR is even higher: 76% (PLR), 131% (SLR), and 338% (MLR).

In Fig. 3, we can inspect the behavior of the predictors during the time course of the 39 data points of the test set in more detail.

The physically measured layer thickness is indicated by a solid black line. The upper and lower tolerance limits (UTL/LTL indicated by the red dashed lines) determine the relative unconditional VM accuracy limit.

In Fig. 3(a), the prediction based on SLR does not follow the fluctuations of the actual measurement in the beginning of the test dataset. Toward the end of the test data, SLR overestimates the measurement. MLR-based prediction follows more or less the general contour of the measurement up to Instance No. 29 of the test data, from which on MLR radically underestimates the layer thickness, however, still echoing the shape of the measurement curve, but with a large offset.

An inspection of the test dataset reveals that just before Instance No. 29 a yearly maintenance of the considered process chamber took place. In addition, it can be noted that during the time period from which the training data was sampled and the time period of five months between training and test set, no maintenance action including a chamber wet clean was performed. In the test set, the FDC variable *count of processed wafers since last chamber wet clean* in the ES variable set (cf., Table I) is far above (before Instance No. 29) and below (after Instance No. 29) its range within the training set. It can be observed that SLR as well as MLR are very sensitive to this variable out of training-set-range condition.

Fig. 3(b) shows the prediction of PLR, RLR, and SVR. PLR and RLR give almost identical results and predict clearly better than MLR. Up to Instance No. 15, the prediction is very close to the actual measurement. From Instance No. 16 to 28, PLR/RLR always overestimate the layer thickness mostly by a large offset. As for MLR, this tendency is reversed from Instances No. 29

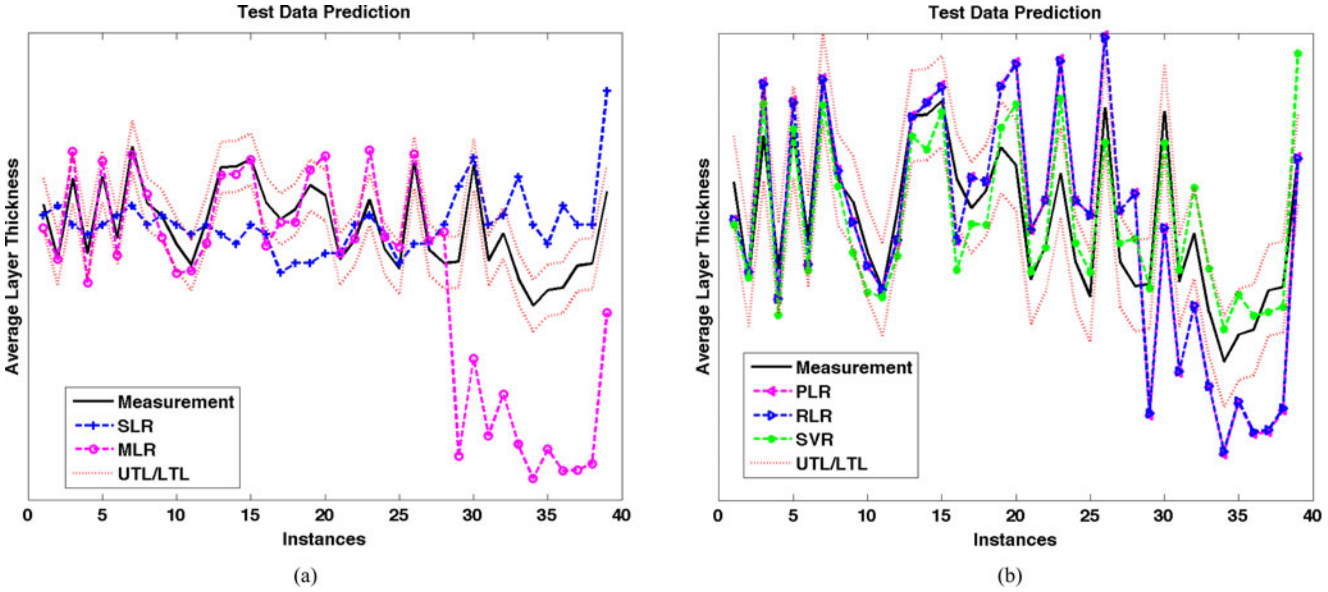

Fig. 3. Prediction of silicon nitride layer thickness on the ES test set, using (a) SLR and MLR, and (b) PLR, RLR, and SVR. The layer thickness measured by optical metrology equipment is indicated by a solid black line. The upper and lower tolerance limits (UTL/LTL) indicated by the red dashed lines determine the relative unconditional VM accuracy limit. The training set is used to optimize the hyperparameters and to train the model. The trained models are then applied to the test set. For the first half of the instances of the test data set for MLR, PLR, and RLR, the performance is relatively good, whereas it significantly degrades from instance No. 29 onwards, which is the first recorded sensor data vector after a yearly maintenance action including wet clean of the process chamber. In contrast to SLR, MLR, PLR, and RLR, the prediction based on SVR follows the actual measurements quite closely and the errors are more equally distributed over the time course of the test data.

to 38, resulting in an underestimation by a large offset. The behavior of the SVR predictor is qualitatively different. The prediction errors of SVR are more evenly distributed across the time course of the test data, and thus, significantly smaller than for the other predictors for the last third of variable instances.

V. Conclusion

Although for the CV on the training set, PLR/RLR perform slightly better than SVR, on the test set, SVR outperforms all linear-regression-based methods by a large margin. This supports the conjecture that all investigated linear-regression-based methods overfit to the training set (e.g., to particular parameter settings), whereas on our test set, SVR seems to generalize better to conditions not explicitly present in the training set. PLR and RLR perform with no significant difference, but clearly better than the unstable MLR. That indicates that previous dimension reduction by PLS estimate improves linear regression, whereas usage of the additional ridge parameter does not improve the prediction perceivably. The univariate SLR is inferior to the multivariate methods PLR/RLR and SVR. These findings converge with [13], where SVR and linear regression in conjunction with previous dimension reduction perform best for VM in an etching process.

Generalizability of the model is particularly challenged if variable values are encountered that are out of the variable range in the training set. This requires extrapolation of the model and can lead to bad performance. It would be advisable to check if the sensor variables exceed the variable range present in the training set. If yes, this particular variable could be removed from the dataset. However, a sufficiently large dataset used for training should minimize the likelihood of such a situation.

The ES variable set gives the best performance, better than the full variable set, the best single variable (*deposition time*), or the best three variables (*deposition time, basic design type,*and *wafer temperature*).

In conclusion, our results indicate that VM can benefit from the usage of robust statistical methods combined with comprehensive process expert knowledge in terms of appropriate selection of the predictor variables.

In future work, a couple of aspects should be investigated. The prediction problem is complicated, because the most predictive variable (*deposition time*) is recalculated by the closed-loop R2R controller; associated predictor variables contain offsets due to maintenance actions, manual adjustments related to SPC or self-regulation of the equipment, without correlated changes of the deposited layer thickness. When the usage of VM is intended to serve as an input to R2R control, the mean squared error is an appropriate evaluation measure. For the purpose of only monitoring the quality of the deposition process with respect to the layer thickness, an alternative would be to formulate a classification problem, distinguishing between three types of behavior: above/within/below specified limits. Physical models have to be designed individually for each new process. On the contrary, the statistical model proposed here does not require deep knowledge about the physical nature of the modeled process. However, to reach high reliability, sufficient data are needed. If this is not the case, the incorporation of some physical process knowledge into the statistical model can compensate to some extent for the lack of data. In order to not rely on process experts to select the predictor variables, methods of automated feature selection [13] will be further investigated. Method candidates include the so-called *filters* that rank individual sensor variables or subsets thereof independently from the regression method. To achieve this, correlation or mutual information can be employed to detect dependence between the variables. An alternative approach are *wrappers* using a regression method, e.g., SVR, to evaluate (subsets of) features. For the latter type of methods, only a relatively small portion of the large number of possible variable combinations should be used for validation to avoid overfitting [20].

References

[1] C. Hollauer, "Modelling of thermal oxidation and stress effects," Ph.D. dissertation, Vienna Univ. Technol., Vienna, Austria, 2007.

[2] M.-H. Hung, T.-H. Lin, F.-T. Cheng, and R.-C. Lin, "A novel virtual metrology scheme for predicting CVD thickness in semiconductor manufacturing," *IEEE/ASME Trans. Mechatronics*, vol. 12, no. 3, pp. 308–316, Jun. 2007.

[3] A. A. Khan, J. Moyne, and D. Tilbury, "Virtual metrology and feedback control for semiconductor manufacturing processes using recursive partial least squares," *J. Process Control*, vol. 18, no. 10, pp. 961–974, 2008.

[4] C.-A. Kao, F.-T. Cheng, and W.-M. Wu, "Preliminary study of run-to-run control utilizing virtual metrology with reliance index," in *Proc. IEEE Conf. Autom. Sci. Eng.*, 2011, pp. 256–261.

[5] A. Grill, "Plasma-deposited diamondlike carbon and related materials," *IBM J. Res. Development*, vol. 43, no. 1–2, pp. 147–161, 1999.

[6] Y.-T. Chen, H.-C. Yang, and F.-T. Cheng, "Multivariate simulation assessment for virtual metrology," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2006, pp. 1048–1053.

[7] A. Ferreira, A. Roussy, and L. Conde, "Virtual metrology models for predicting physical measurement in semiconductor manufacturing," in *Proc. IEEE/SEMI Adv. Semicond. Manuf. Conf.*, 2009, pp. 149–154.

[8] T.-H. Lin, F.-T. Cheng, W.-M. Wu, C.-A. Kao, A.-J. Ye, and F.-C. Chang, "NN-based key-variable selection method for enhancing virtual metrology accuracy," *IEEE Trans. Semicond. Manuf.*, vol. 22, no. 1, pp. 204–211, Feb. 2009.

[9] Y.-T. Huang, F.-T. Cheng, and M.-H. Hung, "Developing a product quality fault detection scheme," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 927–932.

[10] F.-T. Cheng, H.-C. Huang, and C.-A. Kao, "Developing an automatic virtual metrology system," *IEEE Trans. Autom. Sci. Eng.*, vol. 9, no. 1, pp. 181–188, Jan. 2012.

[11] S. Pampuri, A. Schirru, G. Fazio, and G. De Nicolao, "Multilevel lasso applied to virtual metrology in semiconductor manufacturing," in *Proc. Conf. Autom. Sci. Eng.*, 2011, pp. 244–249.

[12] A. Schirru, S. Pampuri, C. De Luca, and G. De Nicolao, "Nonparametric virtual sensors for semiconductor manufacturing—Using information theoretic learning and kernel machines," in *Proc. Int. Conf. Informatics Control, Autom. Robot.*, 2011, [CD-ROM].

[13] P. Kang, H.-J. Lee, S. Cho, D. Kim, J. Park, C.-K. Park, and S. Doh, "A virtual metrology system for semiconductor manufacturing," *Expert Syst. Appl.*, vol. 36, no. 10, pp. 12 554–12 561, 2009.

[14] A. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, 2004.

[15] H. Purwins, A. Nagi, B. Barak, U. Höckele, A. Kyek, B. Lenz, G. Pfeifer, and K. Weinzierl, "Ridge regression for prediction of PECVD silicon nitride layer thickness," in *Proc. IEEE Conf. Autom. Sci. Eng.*, 2011, pp. 387–392.

[16] R. V. Hogg, "An introduction to robust estimation," in *Robustness Statistics.* New York, NY, USA: Academic, 1979, pp. 1–17.

[17] B. Jørgensen and Y. Goegebeur. (2007, Jan.). "Multivariate data analysis and chemometrics," [Online]. Available: http://statmaster.sdu.dk/courses/ST02

[18] C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.

[19] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2005.

[20] I. Guyon, *Feature Extraction: Foundations and Applications*. New York, NY, USA: Springer-Verlag, 2006.

Authors' photographs and biographies not available at the time of publication.